Mod-5 (Statistics)

In statistics, basically we deal with measure of central tendency (Avg. i.e. mean, median or mode) of a given series & its dispersn like std. deviatn, covariance, variance, etc.

// -> Fundamentals:-

(a) In an individual series, if $X \div x_1, x_2, \ldots, x_n$

$$\therefore \text{Mean} = \bar{x} = \frac{\sum x_i}{n}$$

(b) In a discrete series, if $X: x_1, x_2, \ldots, x_n$ & $f: f_1, f_2, \ldots, f_n$, then:

$$\bar{x} = \text{Mean} = \frac{\Sigma f_i x_i}{\Sigma f_i} = \frac{\Sigma f x}{N}$$

(a) In a continuous series, if:

X	f
$x_1 - x_2$	f_1
$x_2 - x_3$	f_2
$x_{n-1} - x_n$	f_{n-1}

$$\therefore \text{Mean} = \bar{x} = \frac{\Sigma f_i m_i}{\Sigma f_i} = \frac{\Sigma f m}{N}$$

where, m = mid-value of class.

(b) Let X: $x_1, x_2, \ldots, x_n$

In this individual series, 1st arrange the data either in ascending or descending order, then:

$$M_d = \text{Median} = \text{Size of } \left(\frac{N+1}{2}\right)^{th} \text{ item.}$$

(a) In a discrete frequency:-

X	f	C.f
x_1	f_1	f_1
x_2	f_2	$f_1 + f_2$
x_n	f_n	Σf_i

For (M_d) i.e. Median, look at the CF (cumulative frequency) column, & set the next higher value at (N/2). The corresponding (X) value is the median.

(b) In a continuous frequency distribut^n

$$M_d = l_1 + \frac{\left(\frac{N}{2}\right) - C}{f} \times (i)$$

where, l_1 = lower limit of median class.

$N = \Sigma f$, f = frequency of median class.

c = CF of the class preceding median class.
i = size of class.

④ The observation having largest frequency is called mode.

④ In continuous frequency distribut^n,

$$M_0 = l_1 + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times (i)$$

where, f_1 = frequency of the modal class
f_0 = frequency of the preceding class
f_2 = frequency of the succeeding class
i = size of the class.
l_1 = lower limit of the modal class

N.B: Mode = 3 Median − 2 Mean (Empirical relat^{n}ship)

// -> Std. deviat^n (σ) or SD: (RMS of deviat^n)

It may be defined as the square root of the arithmetic avg. of the squares of deviat^n taken from the arithmetic avg. of the series.

④ In individual series:-

$$\sigma = \sqrt{\frac{\Sigma (x - \bar{x})^2}{N}}, \quad \bar{x} = \text{Mean}$$

④ In discrete or continuous case:

$$\sigma = \sqrt{\frac{\Sigma f(x - \bar{x})^2}{N}}, \text{ where } N = \Sigma f$$

④ Coefficient of SD:-

$$\text{Coefficient of SD}(\sigma) = \frac{\sigma}{\bar{x}}$$

④ Variance:-

$$V = \sigma^2$$

④ Coefficient of Variance:-

$$CV = \frac{\sigma}{\bar{x}} \times 100$$

// -> Another formula for SD :-

$$SD = \sigma = \sqrt{\frac{\Sigma(x-\bar{x})^2}{N}}$$

$$\Rightarrow \sigma^2 = \frac{\Sigma(x-\bar{x})^2}{N}$$

$$\Rightarrow \sigma^2 = \frac{\Sigma(x^2 - 2x\bar{x} + (\bar{x})^2)}{N}$$

$$= \frac{\Sigma x^2}{N} - 2\bar{x}\frac{\Sigma x}{N} + (\bar{x})^2\frac{\Sigma 1}{N}$$

$$= \frac{\Sigma x^2}{N} - 2\left(\frac{\Sigma x}{N}\right)^2 + \left(\frac{\Sigma x}{N}\right)^2 \cdot \frac{N}{N} \quad \left(\because \bar{x} = \frac{\Sigma x}{N}\right)$$

$$= \frac{\Sigma x^2}{N} - \left(\frac{\Sigma x}{N}\right)^2$$

$$\therefore \sigma = \sqrt{\frac{\Sigma x^2}{N} - \left(\frac{\Sigma x}{N}\right)^2} = \sqrt{\frac{\Sigma x^2}{N} - (\bar{x})^2}$$

$$= \sqrt{\frac{N\Sigma x^2 - (\Sigma x)^2}{N^2}} = \frac{\sqrt{N\Sigma x^2 - (\Sigma x)^2}}{N}$$

Q: Find the Mean & SD of 1st (n) natural nos:

Ans - $N = n$

$$\therefore \text{Mean} = \bar{x} = \frac{\Sigma x}{N} = \frac{1+2+3+\cdots+n}{n}$$

$$= \frac{n(n+1)}{2n} = \frac{n+1}{2}$$

$$\therefore SD = \sigma = \sqrt{\frac{\Sigma x^2}{N} - \left(\frac{\Sigma x}{N}\right)^2}$$

$$= \sqrt{\frac{1^2+2^2+\cdots+n^2}{N} - \left(\frac{n+1}{2}\right)^2}$$

$$= \sqrt{\frac{n(n+1)(2n+1)}{6n} - \frac{(n+1)^2}{4}}$$

$$= \sqrt{\frac{2(n+1)(2n+1) - 3(n+1)^2}{12}}$$

$$= \sqrt{\frac{4n^2+3n+4n+2-3n^2-6n-3}{12}} = \sqrt{\frac{n^2-1}{12}}$$

// -> Correlat^n Analysis :-

- It's used to find the degree of associat^n between 2 variables.
- Correlat^n in statistics refers to the relat^nship between any 2 or more variables like:
  - -> height & wt.
  - -> Rainfall & yield.
  - -> Price & demand.
  - -> Income & expenditure, etc.
- 2 variables are said to be correlated if with the change in the value of 1 variable, there arises a change in the value of another variable.
- The measure of correlat^ is called the correlat^ coefficient or correlat^ index. In 1 fig. it summarises the direct^ & degree of correlat^. It's denoted by (r) & it's value must lie between -1 & 1 (IMP)

// -> Types of correlat^ :-

Based on different kinds of analysis, there're various types of correlat^:

=> Depending upon no. of variables, there're 3 types of correlat^:

- Simple -> when the relat^nship between any 2 variables only is studied it's called simple correlat^.
- Partial -> when the relat^nship between any 2 of 3 or more variables are studied, ignoring the effect of the other related variable, it's a case of partial correlat^
- Multiple -> when the relat^nship between any 3 or more variables is studied at

a time, it's a case of multiple correlat°.

=> Based on direct°: there're 2 types of correlat°:

(a) +ve -> [when the value of both the variables under study move in the same direct°, it's a case of +ve correlat°.]

(b) -ve -> [when both the variables under study move in the opposite direct°, then it's a case of -ve correlat°.]

=> Based on the value of (r): there're 2 types of correlat°:

(a) Perfect -> [when the value of both the variables under study change at a constant ratio, irrespective of the direct°, then it's a case of perfect correlat°.]

(b) Imperfect -> [when the value of both the variables under study change at a different ratio, then it's a case of imperfect correlat°.]

N.B: [When correlat°s are measured mathematically the value of perfect correlat° will be either 1 or -1 & the value of imperfect correlat° will be between -1 & 1.]

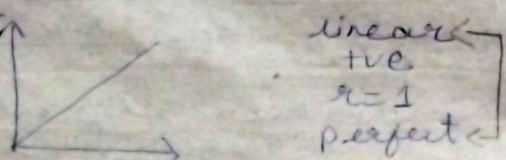
=> Based on graph: there're 2 types of correlat°:

(a) Linear -> [when the data relating to correlat° plotted on a graph rises to a st. line, it's a case of linear correlat°.]

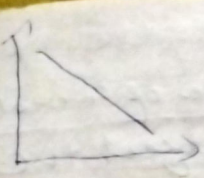
(b) Non-linear -> [when the data of 2 variables plotted on a graph gives out a curve of any direct°, it's a case of curvilinear or non-linear correlat°.]

// -> Types of graph:-

(a)

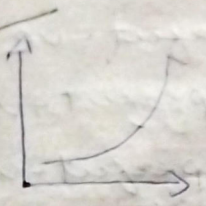


[Linear means Perfect]

(A) 

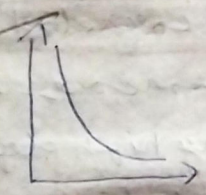
linear
-ve
$r = -1$
perfect

[Linear means Perfect]

(B) 

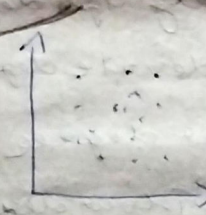
non-linear
+ve
$-1 < r < 1$
imperfect

[Non-linear means Imperfect]

(C) 

non-linear
-ve
$-1 < r < 1$
imperfect

[Non-linear means Imperfect]

(D) 

no-correlat^n.

// -> Karl-Pearson model of correlat^n coefficient:-

The correlat^n coefficient is defined as the ratio of covariance between the 2 variables to the product of their std. deviat^n.

$$\therefore r = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = \frac{\left(\frac{\Sigma xy}{N}\right)}{\sigma_X \sigma_Y}$$

where, (x) is $X - \bar{X}$ & (y) is $Y - \bar{Y}$, where (x) is the deviat^n of the 1st variable from its mean & (y) is the deviat^n of the 2nd variable from its mean, & (N) is the no. of total pairs of variables.

(σ_X) & (σ_Y) are the std. deviat^n of (X) & (Y) respectively.

$$\therefore r = \frac{\Sigma xy}{N\sigma_X \sigma_Y}, \text{ where, } N\sigma_X \sigma_Y = N\sqrt{\frac{\Sigma x^2}{N}}\sqrt{\frac{\Sigma y^2}{N}}$$

$$= \sqrt{\Sigma x^2 \Sigma y^2}$$

$$\Rightarrow r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}} = \frac{\Sigma (X-\bar{X})(Y-\bar{Y})}{\sqrt{\Sigma (X-\bar{X})^2 \Sigma (Y-\bar{Y})^2}}$$

(lengthy formula in terms of calculat^n)

11-> Some facts on (r) :-

Value of (r)	Degree of correlat^n
ⓐ $r = 1$	Perfect +ve
ⓑ $r = -1$	Perfect -ve
ⓒ $0 < r < 1$	Imperfect +ve
ⓓ $-1 < r < 0$	Imperfect -ve
ⓔ $r = 0$	No correlat^n
ⓕ $r > 1$ & $r < -1$	Erroneous result
ⓖ $\lvert r \rvert \geq 0.9$	Very high degree
ⓗ $\lvert r \rvert \geq 0.75$ & < 0.9 $r \in (-0.9, -0.75] \cup [0.75, 0.9)$	Fairly high degree
ⓘ $\lvert r \rvert \geq 0.5$ & ≤ 0.75 $r \in (-0.75, -0.5] \cup [0.5, 0.75)$	Moderate degree
ⓙ $\lvert r \rvert \geq 0.25$ & $\leq \pm 0.5$ $r \in (-0.5, 0.25] \cup [0.25, 0.5)$	Low degree
ⓚ $\lvert r \rvert < 0.25$	very low degree

Q: Find karl-Pearson's correlat^n coefficient for the given data :-

X	8	4	10	2	6
Y	9	11	5	8	7

Ans-

	X	Y	$x = X - \bar{X}$	$y = Y - \bar{Y}$	x^2	y^2	xy
	8	9	2	1	4	1	2
	4	11	-2	3	4	9	-6
	10	5	4	-3	16	9	-12
	2	8	-4	0	16	0	0
	6	7	0	-1	0	1	0
Sum	30	40			40	20	-16

$$\therefore r = \frac{\Sigma xy}{\Sigma x^2 \Sigma y^2} = \frac{-16}{40 \times 20} = -0.57 < -0.5$$

So, there is a moderate degree of -ve imperfect correlat^n between the 2 variables.

11-> Another formula to find (r) :-

we know that $(r) = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$

$$\therefore Cov(X,Y) = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{N}$$

$$= \frac{\Sigma(XY - X\bar{Y} - \bar{X}Y + \bar{X}\bar{Y})}{N}$$

$$= \frac{\Sigma XY}{N} - \frac{\bar{Y}\Sigma X}{N} - \frac{\bar{X}\Sigma Y}{N} + \bar{X}\bar{Y}$$

$$= \frac{\Sigma XY}{N} - \bar{X}\bar{Y} - \bar{X}\bar{Y} + \bar{X}\bar{Y} = \frac{\Sigma XY}{N} - \frac{\Sigma X \Sigma Y}{N^2}$$

$$= \frac{N\Sigma XY - \Sigma X \Sigma Y}{N^2}$$

Also, we know that:

$$\sigma_X = \sqrt{\frac{\Sigma(X-\bar{X})^2}{N}} = \frac{\sqrt{N\Sigma X^2 - (\Sigma X)^2}}{N}$$

$$\sigma_Y = \sqrt{\frac{\Sigma(Y-\bar{Y})^2}{N}} = \frac{\sqrt{N\Sigma Y^2 - (\Sigma Y)^2}}{N}$$

$$\therefore r = \frac{\left(\frac{N\Sigma XY - \Sigma X \Sigma Y}{N^2}\right)}{\frac{\sqrt{N\Sigma X^2 - (\Sigma X)^2}}{N} \times \frac{\sqrt{N\Sigma Y^2 - (\Sigma Y)^2}}{N}}$$

$$\Rightarrow r = \frac{N\Sigma XY - \Sigma X \Sigma Y}{\sqrt{N\Sigma X^2 - (\Sigma X)^2}\sqrt{N\Sigma Y^2 - (\Sigma Y)^2}}$$

(Shortest formula in terms of Calculat^n)

Q: Find the value of (r) for the following data:

X	5	8	10	12	15
Y	1	4	7	10	13

Ans-

X	Y	X^2	Y^2	XY	R_X	R_Y	$(R_X - R_Y)^2$
5	1	25	1	5	1	1	0
8	4	64	16	32	2	2	0
10	7	100	49	70	3	3	0
12	10	144	100	120	4	4	0
15	13	225	169	195	5	5	0

$$\therefore r = \frac{N\Sigma XY - \Sigma X \Sigma Y}{\sqrt{N\Sigma X^2 - (\Sigma X)^2}\sqrt{N\Sigma Y^2 - (\Sigma Y)^2}} = 0.9965 > 0.9$$

So, there is very high degree of +ve imperfect correlat^n.

11-> Spearman's coefficient of rank correlat^n:-

$$r_s = 1 - \frac{6\Sigma D^2}{N^3 - N}$$

where, $\Sigma D^2 = \Sigma (R_X - R_Y)^2$, i.e. sum of the squares of the diff. between the corresponding ranks of the 2 variables.

11-> Rules to give rank:-

(a) In case of ascending order, the smallest rank (1) is assigned to the smallest value of a variable & the subsequent ranks 2, 3, 4, .. are given to the other values, in order of their largeness in size.

(b) Similar is the case for descending order.

(c) When 2 or more values of a variable are found to be identical, each of them is to be assigned with the avg. of their progressive ranks. For ex. suppose 30 is coming after 4th rank thrice i.e. in place of 5th, 6th & 7th posit^n. Then the avg. = (5+6+7)/3 = 6 is assigned to all the 3 identical values & the next value will begin with the rank 8th.

Q: From the following data, find (r_s) & comment on the following result:

Roll No:	1	2	3	4	5	6	7	8	9	10
X:	60	56	25	90	35	14	52	27	54	72
Y:	42	34	56	35	40	50	45	60	58	36

Here, N = 10

Ans-

X	Y	R_X	R_Y	$D = R_X - R_Y$	D^2
60	42	8	5	3	9
56	34	7	1	6	36
25	56	2	8	-6	36
90	35	10	2	8	64
35	40	4	4	0	0
14	50	1	7	-6	36
52	45	5	6	-1	1
27	60	3	10	-7	40
54	58	6	9	-3	9
72	36	9	3	6	36
Sum					276

$$\therefore r_s = 1 - \frac{6\Sigma D^2}{N^3 - N} = -0.673$$

So, moderate degree of -ve imperfect correlat^n is found, i.e. if a student secures more marks in math 1, his marks in math 2 is likely to decrease to a moderate extent.

Q: From the following data, find (r_s) & comment on the following result:

Ans-

X	Y	R_X	R_Y	$D = R_X - R_Y$	D^2
35	71	1	4	-3	9
57	62	3	1.5	1.5	2.25
63	69	8	3	5	25
57	75	3	5	-2	4
61	78	5.5	7	-1.5	2.25
62	79	7	8	-1	1
68	80	9	9	0	0
57	62	3	1.5	1.5	2.25
61	76	5.5	6	-0.5	0.25
80	85	10	10	0	0

$$\therefore r_s = 1 - \frac{6\Sigma D^2}{N^3 - N} = 0.72$$

Here, we've found the moderate degree

of imperfect +ve correlat^ between maths & science, i.e. if a student secures more marks in maths, his marks in science is likely to increase to a moderate extent.

Q-> Regression Analysis:-

- Regress^ is the measure of avg. relat^ship between 2 or more variables.
- It's the study of the nature of relat^ship between the variables so that ~~value of variable for a known~~ 1 may be able to predict the unknown value of 1 variable for a known value of another variable.
- In regress^, 1 variable is considered as an independent variable & another variable is taken as dependent variable.
- With the help of regress^, possible values of the dependent variable are estimated on the basis of the values of the independent variable.

Q-> Purpose of regression:- It's widely used for predict^ & forecasting where its used as substantial overlap with the field of machine learning. Also, in some situat^, it can be used to infere casual relat^ship between independent & dependent variables.

Q-> Types of regress^:-

=> Simple & multiple regress^:

- In simple regress^ analysis, we study only 2 variables at a time, in which, 1 variable is dependent & the other is independent. Ex: the relat^ship between income & expenditure.
- We study 2 or more than 2 variables at a time in multiple regress^ analysis, in which, 1 is dependent & others are independent. Ex: the study of the effect of rain & irrigat^ on yield.

=> Linear & Non-linear regress^:

- When 1 variable changes with other

variable in some fixed ratios, this is called linear regress^. Such type of relat^ship is depicted or plotted on a graph by means of a st. line.

ⓐ When 1 variable varies with other variable in a changing ratio, then it's referred to as curvilinear or non-linear regress^. This relat^ship expressed on a graph paper takes the form of a curve.

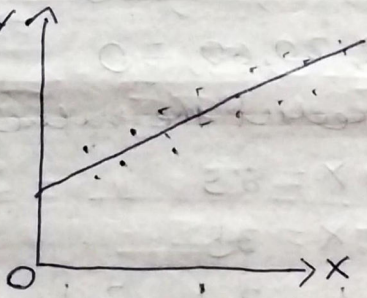
=> Total & Partial regress^:

ⓐ In total regress^, all variables are studied simultaneously for the relat^ship among them.

ⓐ When 2 or more variables are studied for funct^al relat^ship, but at a time, the relat^ship between only 2 variable is studied & other variables are held constant are called partial regress^.

// -> Regress^ lines:-

ⓐ It's a line that most closely fits the data according to some specific mathematical criterium.



ⓐ For 2 variables, we've 2 regress^ lines

Y on X	X on Y
ⓐ $Y = a + bX$	ⓐ $X = c + dY$
where, Y = dependent X = independent a & b are constants	where, Y = independent X = dependent c & d are constants

// -> Calculat^ of regress^ line equ^:-

=> By least square method -> It's also known as regress^ equn. using normal

equns. Under this method, the sum of the squares of deviat^n between the given value of a variable & its estimated value given by the concerned line of the regress^n is the least or minimum possible. So, we can draw the lines of best fit as the lines of regress^n under this method.

1-> Process:

(a) To find the regress^n line of Y on X i.e. $Y = a + bX \rightarrow$

$\therefore Y = a + bX \Rightarrow \Sigma Y = aN + b\Sigma X$ ---①

$\Sigma XY = a\Sigma X + b\Sigma X^2$ ---②

Solve & find the value of (a) & (b).

(b) To find the regress^n line of X on Y i.e. $X = c + dY \rightarrow$

$\therefore X = c + dY \Rightarrow \Sigma X = cN + d\Sigma Y$ ---①

$\Sigma XY = c\Sigma Y + d\Sigma Y^2$ ---②

Solve & find the value of (c) & (d).

Q: Draw the 2 regress^n lines for the following data using least square method:

X: 5, 10, 15, 20, 25

Y: 20, 40, 30, 60, 50 $\Rightarrow N = 5$

& then, predict the value of:

(a) X, when Y = 85

(b) Y, when X = 35

Ans-

X	Y	X^2	Y^2	XY
5	20	25	400	100
10	40	100	1600	400
15	30	225	900	450
20	60	400	3600	1200
25	50	625	2500	1250
Sum: 75	200	1375	9000	3400

(a) The regress^n line X on Y => $X = c + dY$

$\therefore \Sigma X = cN + d\Sigma Y$ & $\Sigma XY = c\Sigma Y + d\Sigma Y^2$

$\Rightarrow 75 = 5c + 200d$

$3400 = 200c + 9000d$

$\Rightarrow c = -1$ & $d = 0.4 \Rightarrow \boxed{X = -1 + 0.4Y}$

-OR-

$$\boxed{d = \frac{N\Sigma XY - \Sigma X \Sigma Y}{N\Sigma X^2 - (\Sigma X)^2} = 0.4}$$

$$\boxed{c = \frac{\Sigma X - d\Sigma Y}{N} = -1}$$

$\therefore \boxed{X = -1 + 0.4Y}$

(b) The regress^n line Y on X $\Rightarrow Y = a + bX$

$\therefore \Sigma Y = aN + b\Sigma X$ & $\Sigma XY = a\Sigma X + b\Sigma X^2$

$\Rightarrow 200 = 5a + 75b$

$3400 = 75a + 1375b$

$\Rightarrow a = 16$ & $b = 1.6 \Rightarrow \boxed{Y = 16 + 1.6X}$

-OR-

$$\boxed{b = \frac{N\Sigma XY - \Sigma X \Sigma Y}{N\Sigma X^2 - (\Sigma X)^2} = 1.6}$$

$$\boxed{a = \frac{\Sigma Y - b\Sigma X}{N} = 16}$$

$\therefore \boxed{Y = 16 + 1.6X}$

(c) Y on X:

For $X = 5$ & $25 \Rightarrow Y = 24$ & 56

$\therefore$ when $X = 35 \Rightarrow Y = 72$

(d) X on Y:

For $Y = 20$ & $60 \Rightarrow X = 7$ & 23

$\therefore$ when $Y = 85 \Rightarrow X = 33$

11 → To find regress^n lines by method of deviat^n from actual mean:-

Using this method, 2 regress^n equns. are modified as:

① Regression line of X on Y is:

$$X - \bar{X} = b_{xy}(Y - \bar{Y}) \quad \text{---(i)}$$

where, $\bar{X}$ = mean of X

$\bar{Y}$ = mean of Y

b_{xy} = Regression coefficient of X on Y

$$\Rightarrow b_{xy} = r\frac{\sigma_x}{\sigma_y} \quad \text{---(ii)}$$

where, r = correlation coefficient.

σ_x = SD of X

σ_y = SD of Y

② Regression line of Y on X is:

$$Y - \bar{Y} = b_{yx}(X - \bar{X}) \quad \text{---(iii)}$$

where, b_{yx} = Regression coefficient of Y on X

$$\Rightarrow b_{yx} = r\frac{\sigma_y}{\sigma_x} \quad \text{---(iv)}$$

∴ From (ii) & (iv), we've:

$$b_{xy} \cdot b_{yx} = r^2 \frac{\sigma_x}{\sigma_y} \cdot \frac{\sigma_y}{\sigma_x} \Rightarrow r = \sqrt{b_{xy} \cdot b_{yx}}$$

// → Formula to calculate b_{xy} & b_{yx}:-

① $b_{xy} = r\frac{\sigma_x}{\sigma_y} = \frac{\Sigma xy}{N\sigma_x\sigma_y} \cdot \frac{\sigma_x}{\sigma_y} = \frac{\Sigma xy}{N\sigma_y^2}$

$$= \frac{\text{Cov}(X,Y)}{(\sigma_y)^2} = \frac{\frac{\Sigma xy}{N}}{\frac{\Sigma y^2}{N}} = \frac{\Sigma xy}{\Sigma y^2}$$

$$\therefore b_{xy} = \frac{\Sigma xy}{\Sigma y^2} = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{\Sigma(Y-\bar{Y})^2}$$

② $b_{yx} = r\frac{\sigma_y}{\sigma_x} = \frac{\Sigma xy}{N\sigma_x\sigma_y} \cdot \frac{\sigma_y}{\sigma_x} = \frac{\Sigma xy}{N\sigma_x^2}$

$$= \frac{\text{Cov}(X,Y)}{(\sigma_x)^2} = \frac{\frac{\Sigma xy}{N}}{\frac{\Sigma x^2}{N}} = \frac{\Sigma xy}{\Sigma x^2}$$

$$\therefore b_{yx} = \frac{\Sigma xy}{\Sigma x^2} = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{\Sigma(X-\bar{X})^2}$$

(i) $b_{xy} = \frac{cov(X,Y)}{(\sigma_y)^2} = \frac{N\Sigma XY - \Sigma X \Sigma Y}{N\Sigma Y^2 - (\Sigma Y)^2}$

(ii) $b_{yx} = \frac{cov(X,Y)}{(\sigma_x)^2} = \frac{N\Sigma XY - \Sigma X \Sigma Y}{N\Sigma X^2 - (\Sigma X)^2}$

Q: Using the ~~least~~ method of deviat^n from actual mean, find:

(a) the 2 regress^n lines.
(b) the correlat^n coefficient.
(c) the most probable value of Y, when X = 30.

Ans-

X	Y	X^2	Y^2	XY
25	43	625	1849	1075
28	46	784	2116	1288
35	49	1225	2401	1715
32	41	1024	1681	1312
31	36	961	1296	1116
36	32	1296	1024	1152
29	31	841	961	899
38	30	1444	900	1140
34	33	1156	1089	1122
32	39	1024	1521	1248
SUM 320	380	10380	14838	12067

(a) Y on X:

$$Y - \bar{Y} = b_{yx}(X - \bar{X})$$

$$\therefore b_{yx} = \frac{N\Sigma XY - \Sigma X \Sigma Y}{N\Sigma X^2 - (\Sigma X)^2} = -0.6643$$

$$\therefore Y - \frac{380}{10} = -0.6643\left(X - \frac{320}{10}\right)$$

$$\Rightarrow \boxed{Y = 59.257 - 0.6643X}$$

(b) X on Y:

$$X - \bar{X} = b_{xy}(Y - \bar{Y})$$

$$b_{xy} = \frac{N\Sigma XY - \Sigma X \Sigma Y}{N\Sigma Y^2 - (\Sigma Y)^2} = -0.2337$$

$\therefore X - \frac{320}{10} = -0.2337\left(Y - \frac{380}{10}\right)$

$\Rightarrow \boxed{X = 40.881 - 0.2337Y}$

$\therefore r = \sqrt{b_{xy} \cdot b_{yx}} = -0.394$

N·B: Std deviat^n can never be -ve

(a) When $X = 30 \Rightarrow Y = 39.3286$

N·B: (a) Regress^n line of X on Y $\Rightarrow$

$\boxed{X - \bar{X} = b_{xy}(Y - \bar{Y})}$

$\Rightarrow X = (\bar{X} - b_{xy}\bar{Y}) + (b_{xy})Y \Rightarrow X = C + dY$

$\therefore \boxed{C = \bar{X} - b_{xy}\bar{Y}}$ & $\boxed{d = b_{xy}}$

(b) Regress^n line of Y on X $\Rightarrow$

$\boxed{Y - \bar{Y} = b_{yx}(X - \bar{X})}$

$\Rightarrow Y = (\bar{Y} - b_{yx}\bar{X}) + (b_{yx})X \Rightarrow Y = a + bX$

$\therefore \boxed{a = \bar{Y} - b_{yx}\bar{X}}$ & $\boxed{b = b_{yx}}$

Q: Given, variance of $X = 25$

Regress^n of X on Y $\Rightarrow 5X - Y = 22$

Regress^n of Y on X $\Rightarrow 6X - 45Y = 24$

(a) Find the mean value of X & Y

(b) Coefficient of correlat^n between X & Y

(c) Std. deviat^n of Y.

Ans- (a) Regress^n of X on Y:

$\boxed{5X - Y = 22} \Rightarrow X = \frac{22}{5} + \frac{Y}{5} \Rightarrow X = C + dY$

$\therefore C = \frac{22}{5}$ & $d = \frac{1}{5}$

But, $\bar{X} - b_{xy}\bar{Y} = C$ & $b_{xy} = d$

$\Rightarrow \boxed{5\bar{X} - \bar{Y} = 22}$ - - - (i)

(b) Regress^n of Y on X:

$6X - 45Y = 24 \Rightarrow Y = -\frac{24}{45} + \frac{6X}{45} \Rightarrow Y = a + bX$

$\therefore a = -\frac{24}{45}$ & $b = \frac{6}{45}$

But, $\bar{Y} - b_{yx}\bar{X} = a$ & $b_{yx} = b$

$6\bar{X} - 45\bar{Y} = 24$

$\Rightarrow \boxed{15\bar{Y} - 2\bar{X} = -8}$ ... (ii)

Now, solving equns (i) & (ii), we've:

$\bar{X} = 4.41$ & $\bar{Y} = 0.055$

Again, $r = \sqrt{b_{xy} \cdot b_{yx}} = \sqrt{(1/5) \times (2/15)} = 0.163$

Now, $Var(X) = (\sigma_X)^2 = 25 \Rightarrow \sigma_X = 5$

$\therefore b_{xy} = r\frac{\sigma_x}{\sigma_y} \Rightarrow \sigma_y = \frac{\sigma_x \cdot r}{b_{xy}} = 4.075$

// -> Sampling & statistical inference:-

// -> Populat^n & Sample ->

- The set of obs. related to a phenomenon under investigat^n is called populat^n.
- A finite subset or a fraut^n of statistical individual in a populat^n is called sampling.
- The no. of individuals in a sample is called the sample size.

// -> Sampling -> The method or procedure of drawing a sample from the populat^n is called sampling. Basically, sampling is of 2 types:

- Unbiased/Random/Probability sampling.
- Biased/Fixed/Non-Probability Sampling.

// -> Parameter & Statistics:-

(IMP) Parameter is a characteristics of populat^n & statistics is a characteristic of sampling.

Parameter	Statistics
Populat^n mean (μ)	Sample mean ($\bar{x}$)
Populat^n propor^n (P)	Sample propor^n ($\bar{P}$)
SD (σ)	SD (s)

- SD(σ)
- Variance(σ^2)

- SD(s)
- Variance(s^2)
- Sample std. error

$$S_{\bar{x}} = \frac{s}{\sqrt{n}}$$

N.B: - A sample statistic is used to estimate populat^n parameter (IMP)
- n = sample size
- N = Populat^n size

// -> Statistical inference: - The process of drawing inference about populat^n on the basis of sample data is called statistical inference. There're 2 types of statistical inference:
- Estimat^n.
- Test of Hypothesis.

// -> Estimat^n -> A process in which we obtain the values of unknown populat^n parameters with the help of sample data is called estimat^n.

// -> Estimator -> It's a ~~state~~ rule or formula or funct^n that tells how to calculate an estimate.

// -> Estimate -> It is a rule or formula or funct^n that tells how to calculate an estimate. It's the observed numeric value of the estimator

// -> Types of estimat^n :-> There're 2 types of estimat^n:
- Pt. estimat^n.
- Interval estimat^n.

// -> Pt. estimat^n -> When an estimate for the unknown populat^n parameter is expressed by a single value, the process is called pt. estimat^n.

// -> Interval estimat^n -> When an estimate for the unknown populat^n parameter is expressed by a range of values, within which, the populat^n parameter is ~~expected~~

expected to occur is called an interval estimate, & the process is called interval estimat^n.

//-> Pt. Estimat^n :- There are 2 main methods of pt. estimat^n:

(a) Max. Likelihood Estimat^n (MLE) method
(b) Method of Moments (MOM)

//-> MLE method -> Let the funct^n (PDF or PMF) be $f(x,\theta)$, where (θ) is a parameter.

S-1: Consider the likelihood funct^n as:

$$L = f(x_1,\theta)\cdot f(x_2,\theta)\cdots f(x_n,\theta)$$

$$\Rightarrow \boxed{L = \prod_{i=1}^{n} f(x_n,\theta)}$$

S-2: Take natural log on both sides:

$$\boxed{\ln(L) = \ln\left(\prod_{i=1}^{n} f(x_n,\theta)\right)}$$

S-3: So, for max. value, equate:

$$\boxed{\frac{\partial}{\partial\theta}(\ln(L)) = 0}$$

to get the value of (θ) [i.e. populat^n parameter in terms of sample parameter]

=> For Binomial distribut^n ->

$$\boxed{f(x) = {}^{n}C_{x}\cdot p^{x}\cdot(1-p)^{n-x}}$$

where, n = no. of trials
p = probability of success
1-p = probability of failure

x = no. of successes

Here, to estimate the populat^n parameter, we can proceed with the parameter (p).

=> For Poisson's distribut^n ->

$$f(x) = \frac{e^{-\mu} \cdot \mu^x}{x!}$$

where, μ = mean (Populat^n)

x = no. of successes

=> For Normal distribut^n ->

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Populat^n Populat^n

where, μ = mean ; σ = S.D ; σ^2 = Variance

Q: Find the MLE of (θ) in the density:

$$f(x) = \begin{cases} \theta e^{-\theta x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$ -> PDF of ERV

Ans- Let's consider the likelihood funct^n:

$$L = \prod_{i=1}^{\hat{n}} f(x_i, \theta) = \theta e^{-\theta x_1} \cdot \theta e^{-\theta x_2} \cdots \theta e^{-\theta x_n}$$

$$\Rightarrow L = \theta^{\hat{n}} e^{-\theta \sum_{i=1}^{\hat{n}} x_i}$$

Take (ln) on both sides. Then, we've:

$$\Rightarrow \ln L = \ln\left(\theta^{\hat{n}} e^{-\theta \sum_{i=1}^{\hat{n}} x_i}\right) = n \ln \theta - \theta \sum_{i=1}^{\hat{n}} x_i \ln e$$

$$= n \ln \theta - \theta \sum_{i=1}^{\hat{n}} x_i$$

Now, to get the max. value, equate the partial derivative of ln(L) w.r.t (θ) with zero.

$$\therefore \frac{\partial}{\partial \theta}(\ln L) = \frac{\partial}{\partial \theta}\left[n \ln \theta - \theta \sum_{i=1}^{\hat{n}} x_i\right]$$

$$= n \cdot \frac{1}{\theta} - \sum_{i=1}^{n} x_i \cdot \frac{\partial \theta}{\partial \theta} = \frac{n}{\theta} - \sum_{i=1}^{n} x_i = 0$$

$$\Rightarrow \frac{n}{\theta} = \sum_{i=1}^{n} x_i \quad \Rightarrow \frac{1}{\theta} = \frac{\sum_{i=1}^{n} x_i}{n} \Rightarrow \frac{1}{\theta} = \bar{x}$$

$$\Rightarrow \boxed{\theta = \frac{1}{\bar{x}}}$$ So, using MLE method, (θ) for the above density funct^ is $\frac{1}{\mu}$

$$\Rightarrow \boxed{\theta = \frac{1}{\mu}}$$ → For ERV

Q: Find the MLE for Poiss^'s distribut^.

Ans- Here, PDF is: $f(x) = \frac{e^{-\mu} \cdot \mu^{x}}{x!}$

$$\therefore L = \prod_{i=1}^{n} f(x_i, \mu) = \prod_{i=1}^{n} \frac{e^{-\mu} \mu^{x_i}}{x_i!}$$

$$\Rightarrow L = \frac{e^{-n\mu} \cdot \mu^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!}$$

$$\Rightarrow \ln L = -n\mu + \sum_{i=1}^{n} x_i \cdot \ln \mu - \ln \left(\prod_{i=1}^{n} x_i! \right)$$

$$\Rightarrow \frac{\partial}{\partial \mu} (\ln L) = -n + \sum_{i=1}^{n} x_i \cdot \frac{1}{\mu} = 0$$

$$\Rightarrow n = \frac{\sum_{i=1}^{n} x_i}{\mu} \Rightarrow \mu = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x} \quad \therefore \boxed{\mu = \bar{x}}$$

For Poiss^'s Random Variable

Q: Estimate the value of mean of normal distribut^ using MLE method.

Ans- Here, PDF is: $f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

$$\therefore L = \prod_{i=1}^{n} f(x_i, \mu) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \cdot e^{-\frac{1}{2\sigma^2}\left[(x_1-\mu)^2 + \cdots + (x_n-\mu)^2\right]}$$

$$\Rightarrow L = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \cdot e^{-\frac{1}{2\sigma^2}\cdot\sum_{i=1}^{n}(x_i-\mu)^2}$$

$$\Rightarrow \ln L = n\left[\ln 1 - \ln \sigma\sqrt{2\pi}\right] - \frac{1}{2\sigma^2}\cdot\sum_{i=1}^{n}(x_i-\mu)^2$$

$$\Rightarrow \frac{\partial}{\partial\mu}(\ln L) = \frac{\partial}{\partial\mu}(n\ln 1 - n\ln\sigma\sqrt{2\pi}) - \frac{1}{2\sigma^2}\cdot\sum_{i=1}^{n}\frac{\partial}{\partial\mu}(x_i-\mu)^2$$

$$= -\frac{1}{2\sigma^2}\cdot\sum_{i=1}^{n}\cdot 2(x_i-\mu)(0-1) = 0$$

$$\Rightarrow \sum_{i=1}^{n} x_i - \mu\sum_{i=1}^{n} 1 = 0 \Rightarrow \sum_{i=1}^{n} x_i - n\mu = 0$$

$$\Rightarrow \mu = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x} \quad \therefore \boxed{\mu = \bar{x}} \Rightarrow \text{For NRV}$$

11-> Method of Moments:-

(a) In statistics, MOM is a method of estimat^n of populat^n parameter.

(b) It starts by expressing the populat^n moments i.e. the expected values of powers of the random variables under considerat^n as funct^{n}s of the parameters of interest.

(c) Those expressi^{n}s are then set equal to the sample moments. The no. of such eqn^s is same as the no. of parameters to be

estimated.

④ These equns are then solved for the parameters of interest & the solut'ns are the estimates of those parameters.

Q: Let (X) be a NRV i.e. $X \sim N(\mu, \sigma^2)$. Find moment estimate of μ & σ^2. (Parameters of interest)

Ans: Let $x_1, x_2, \ldots, x_n$ be a random sample of size (n) from $N(\mu, \sigma^2)$ populat^n.

$\therefore f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ is a NRV.

Here, we've 2 unknown parameters. So we've to solve 2 equns, given by:

$$\mu_1 = m_1 \text{ -- (i)} \quad \& \quad \mu_2 = m_2 \text{ -- (ii)}$$

where, μ_1, μ_2 = Populat^n moments
m_1, m_2 = Sample moments

Now, $\mu_1 = \boxed{\frac{\Sigma x}{n} = \bar{x}}$ -- (iii) $\left(\because m_1 = \frac{\Sigma x}{n}\right)$ & $(\mu_1 = \bar{x})$

& $\mu_2 = \frac{\Sigma x^2}{n}$ -- (iv) $\left(\because m_2 = \frac{\Sigma x^2}{n}\right)$

But,

(i) $\mu_1 = E[X] = \int_{-\infty}^{\infty} x f(x)\, dx = \mu$

(ii) $\mu_2 = E[X^2] = \int_{-\infty}^{\infty} x^2 f(x)\, dx = \sigma^2 + \mu^2$

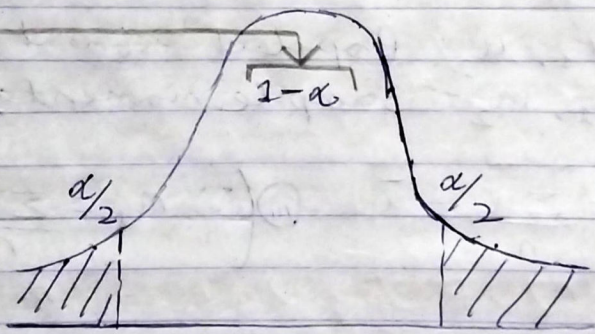
$\therefore$ Putting the value of μ_1 & μ_2 in equn. (iii) & (iv), we've:

$\boxed{\mu = \bar{x}}$ & $\sigma^2 + \mu^2 = \frac{\Sigma x^2}{n} \Rightarrow \sigma^2 = \frac{\Sigma x^2}{n} - \mu^2$

But, $\mu = \bar{x} \Rightarrow \boxed{\sigma^2 = \frac{\Sigma x^2}{n} - (\bar{x})^2}$

// -> Interval estimat^ -> When an estimate for the unknown populat^ parameter is expressed by a range of values within which, the populat^ parameter is estimated to occur is called interval estimate.

// -> Confidence interval -> A range of values within which the populat^ parameter is expected to occur is called confidence interval & is denoted by $1-\alpha$, where (α) is the level of significance.



// -> Confidence limits -> The 2 end pts. of the confidence interval are called confidence limits.

// -> Confidence interval for (μ) :-

(1) When $n \geq 30$ & (σ) is known, then the confidence interval is given by:

$$\bar{x} \pm Z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

where, $\bar{x}$ = sample mean
σ = populat^n S.D
n = sample size
$Z_{\frac{\alpha}{2}}$ = critical value of (z) distribut^n.

(a) when $n \geq 30$ & (σ) is unknown, then the confidence interval is given by:

$$\bar{x} \pm Z_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}$$

where, s = sample S.D

(b) when $n < 30$ & (σ) is unknown, then the confidence interval is given by:

$$\bar{x} \pm Z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

(c) when $n < 30$ & (σ) is unknown, then the confidence interval is given by:

$$\bar{x} \pm t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}$$

where, $t_{\frac{\alpha}{2}}$ = critical value of (t) distribut^n.

Q: An electrical firm manufactures light bulbs that've a length of life with mean(μ) & a S.D of 40 hrs. If a sample of 100 bulbs has an avg. life of 780 hrs, find the 95% confidence interval for the populat^n mean of all bulbs produced by this firm.

Ans- Given, $n = 100 \geq 30$; $\sigma = 40$ hrs; $\bar{x} = 780$;
$1 - \alpha = 0.95 \Rightarrow \alpha = 0.05$

$\therefore$ $n = 100$ & σ is known, so: $\bar{x} \pm Z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$

But, $Z_{\frac{\alpha}{2}} = 1.960$

$\Rightarrow 772.16 \leq \mu \leq 787.84$

Q: The ht. of a random sample of 50 college students showed a mean of 174.5 cm & a S.D of 6.9 cm. construct a 98% confidence interval for the mean ht. of all college students.

Ans- Given, $n = 50 \geq 30$; $S = 6.9$; $\bar{x} = 174.5$;
$1 - \alpha = 0.98 \Rightarrow \alpha = 0.02$

$\because n = 50$ & (σ) is unknown, so: $\bar{x} \pm Z_{\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}}$

But, $Z_{\frac{\alpha}{2}} = 2.3263$

$\Rightarrow 172.23 \leq \mu \leq 176.77$

//-> Test of Hypothesis or Significance:- A very important aspect of sample theory is the study of the Test of Hypothesis.

//-> Hypothesis Testing:-

=> Null Hypothesis (H_0):

- A definite stt. about a populat^n parameter which is usually a hypothesis of no difference is called a null hypothesis, denoted by (μ_0).

- Also, the assumpt^s we use to test is called null hypothesis.

- Also, null hypothesis is the hypothesis which is tested for possible reject^n under the assumpt^n that it's true.

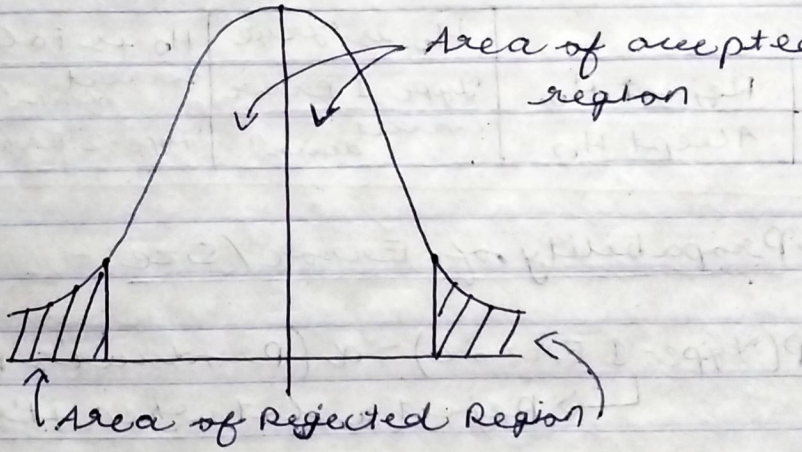
=> Alternative Hypothesis (H_1): Whenever we reject null hypothesis, the conclusn we do accept is called the alternative hypothesis.

* Also, we can say any hypothesis, which is complementary to the null hypothesis is called the alternative hypothesis, denoted by H_1.

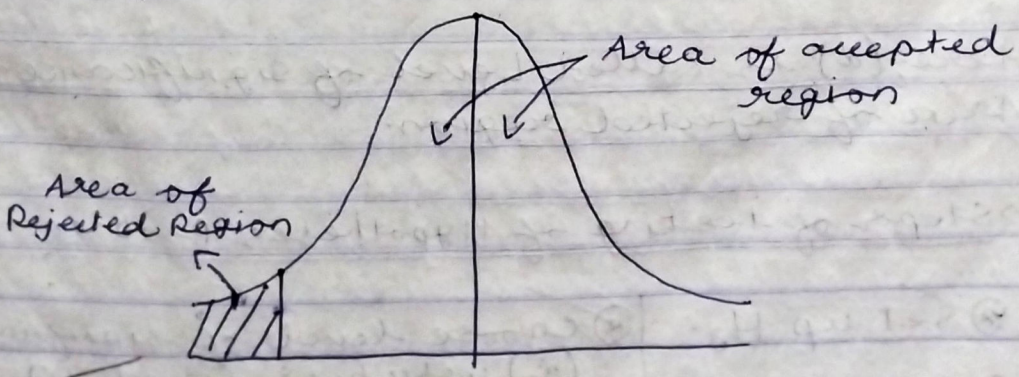
∴

Ex: If we want to test the null hypothesis that the populatn has a specified mean (μ_0). i.e. $H_0: \mu = \mu_0$, then the alternative hypothesis is:

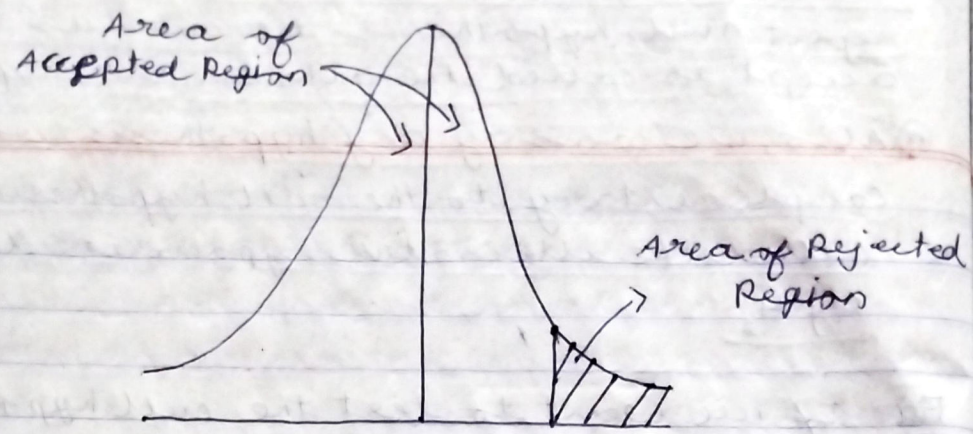
* $H_1: \mu \neq \mu_0$ (i.e. $\mu > \mu_0$ or $\mu < \mu_0$) is known as 2 tailed alternative.



* $H_1: \mu < \mu_0$ is known as left tailed ~~region~~. alternative.



* $H_1: \mu > \mu_0$ is known as right tailed alternative.



// → Error in Sampling :-

① Type-1 Error → Reject H_0 when it's True
② Type-2 Error → Accept H_0 when it's False
i.e. accept H_0 when H_1 is True

	H_0 is True	H_0 is False
Reject H_0	Type-1 Error	correct decision
Accept H_0	correct decision	Type-2 Error

// → Probability of Error / Size :-

① P(Type-1 Error) = α (Producer's Risk)
↳ Reject a lot when it's Good.

② P(Type-2 Error) = β (Consumer's Risk)
↳ Accept a lot when it's Bad.

Here, (α) is called Level of significance or Area of rejected region.

// → Steps of testing of Hypothesis Problem :-

① Set up H_0.
② Set up H_1.
③ Choose level of significance (α) which is fixed in advance. Take 5% when (α) is unknown.

④ Decide the type of the test is required (Z or t or F or χ^2).

⑤ Compute the test statistic. (It can be calculated from normal data using formula)

⑥ Find the critical value i.e. $Z_{\alpha/2}$ from the table.

⑦ Decision

⑧ Conclusion

→ If $|z| < Z_{\alpha/2}$ i.e. it isn't significant. So null hypothesis is accepted.

→ If $|z| > Z_{\alpha/2}$ i.e. it is significant, & hence null hypothesis is rejected at level of significance (α).

II-> Test of Significance of mean :-

① Set up $H_0 : \mu = \mu_0$

② Set up $H_1 : \mu \neq \mu_0$

$H_1 : \mu < \mu_0$

$H_1 : \mu > \mu_0$

③ Choose (α).

④ Decide whether (t) test or (Z) test is applicable or not.

⑤ Find the value of the test statistic (t or Z)

$$\therefore t \text{ or } Z = \frac{\bar{x} - \mu_0}{\left(\frac{\sigma}{\sqrt{n}}\right)} \text{ or } \frac{\bar{x} - \mu_0}{\left(\frac{s}{\sqrt{n}}\right)}$$

⑥ Get the critical value from the table (t or Z).

⑦ Conclusion.

Q: The mean lifetime of a sample of 100 tube lights produced by a company is computed to be 1510 hrs with a S.D of 120 hrs. The company claims that the avg. life of the tube produced ~~by 5% level of sign~~ by the company is atleast 1600 hrs. Using 5% level of significance, test the claim of manufacture.

Ans: Given, $n = 100$, $\bar{x} = 1510$, $S = 120$, $\mu_0 = 1600$
$\alpha = 0.05$

$\therefore H_0 : \mu = \mu_0 = 1600$
$\Rightarrow H_1 : \mu \neq \mu_0 = 1600$

Here, $n \geq 30$ & σ is unknown. So, Z-test is applicable.

$$\therefore Z = \frac{\bar{x} - \mu_0}{\left(\frac{S}{\sqrt{n}}\right)} = -7.5 \quad \& \quad Z_{\frac{\alpha}{2}} = 1.960$$

$\therefore |Z| = 7.5 > Z_{\frac{\alpha}{2}} = 1.960$

So, here, the null hypothesis is rejected. This means that the company is making a false claim.

Q: 10 individuals are chosen at random from a populat^n & their heights are found to be 63, 63, 66, 67, 68, 69, 70, 70, 71, 71 inches. Test the

hypothesis that the mean height of the population is 66 inches.

Ans- Given, $n = 10$, $\mu_0 = 66$

Here, $\bar{x} = \frac{63+63+66+67+68+69+70+71+73}{10}$

$= 67.8$

Again, $S = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}} = 3.011$

Let, $\alpha = 0.05$. Here, $n \leq 30$ & σ is unknown

So, t-test is applicable.

$\therefore t = \frac{\bar{x} - \mu_0}{\left(\frac{S}{\sqrt{n}}\right)} = 1.890$ & $t_{\frac{\alpha}{2}} = 2.262$

$\therefore |t| = 1.890 < t_{\frac{\alpha}{2}} = 2.262$

So, here the null hypothesis is accepted.